On the Impact of YADAS at LANL

Todd L. Graves (tgraves@lanl.gov)
Statistical Sciences, Los Alamos National Laboratory

August 20, 2008

1 Introduction

YADAS allows CCS-6 statisticians to do quick first cuts at analyses which are defensible on projects with short deadlines. The short time allowed eliminates the possibility of writing your own code, e.g., in C. Moreover, if the model changes, you might have to start writing new code from scratch. On the other hand, model changes often means a trivial change in YADAS. Consequently, the statistician can focus on developing the most faithful model for the data at hand and leave the implementation to YADAS. Also, YADAS has some automatic tuning or the ability to change the sampling algorithm to improve convergence. Note that the commercially available WinBUGS often cannot be used to implement the analyses required by LANL.

The YADAS web site (yadas.lanl.gov) features links to download the software, an introduction to the software including a tutorial, examples, and links to related paper. Examples are "YADAS/YATRAN tutorial, version 2008.2", Todd L Graves, LA-UR-08-05485.

2 List of Models Implemented in YADAS

In theory, YADAS can handle any analysis. By itself this statement is meaningless, since it doesn't imply that it is easy to do any analysis. However, we have recorded many successes. The following is a partial list of statistical models that have been implemented in YADAS.

1. system reliability

   - Graves, T.L. (2006). The YADAS Reliability Package. Los Alamos National Laboratory Report LA-UR-06-7739. *This writeup documents how to perform several of the reliability analyses discussed below in YADAS.*
   - V. Johnson, T. Graves, M. Hamada and C.S. Reese. (2003). A hierarchical model for estimating the reliability of complex systems (with discussion). In *Bayesian Statistics 7*, Oxford University Press, Eds. J.M. Bernardo, M.J. Bayarri, J. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West, 199-213. *This is the first paper showing how to combine multi-level binomial data and information for system reliability assessment.*
   - T. Graves and M. Hamada, "Assessing system reliability and allocating resources: a bayesian approach that integrates multi-level data" LA-UR-08-05484; T. Graves and M. Hamada. (2004). Combining multi-level data to assess system reliability and allocate resources optimally: Bayesian methods and computation; LA-UR-04-7157. Submitted to *Journal of Quality Technology. This paper considers system reliability assessment using multi-level binomial data and information for two real systems and shows how to do resource allocation on a simpler system.*
   - T. Graves and M. Hamada. Bayesian methods for assessing system reliability: models and computation (2005). In *Modern Statistical and Mathematical Methods in Reliability*, World Scientific Publishing Company, Eds. A. Wilson, N. Limnios, S. Keller-McNulty, Y. Armijo, 41-54. *This paper shows how to combine multi-level diverse data (system binomial, component logistic regression, component lifetime, component degradation) and information for system reliability assessment. This paper also considers resource allocation for multi-level binomial data.*

- A.G. Wilson, T. Graves, M. Hamada and C.S. Reese. (2006). Advances in data combination and collection for system reliability assessment. *Statistical Science* 21:514-531. *This paper considers hierarchical models, the combination of multi-level diverse data and information, alternate representations (Bayes nets and flowgraph models) and priors, and resource allocation for system reliability assessment. For resource allocation, this paper allows for a discrepancy between the reliability based on system level data alone and the reliability obtained from component level data assuming the system reliability block diagram holds exactly for independent components.*
- Anderson-Cook, C.A., T. Graves, N. Hengartner, R. Klamann, A.Koehler, A.G. Wilson, G. Anderson, and G. Lopez. (2005). Reliability Modeling Using Both System Test and Quality Assurance Data. Los Alamos National Laboratory Report LA-UR-04-2252. *This paper analyzes a system with several components, with partially informative binary system test data, with some of the components having one or more specification tests, and with one or more covariates. The same model has been used in a large simulation study by Vander Wiel and Lawrence investigating the value of different types of information and how to construct vague priors appropriately. The simulation study requires MCMC software that runs on simulated data with no human input.*
- C. Anderson-Cook, T. Graves, M. Hamada, N. Hengartner, V. Johnson, C.S. Reese and A.G. Wilson. (2007). Bayesian stockpile reliability methodology for complex systems with application to a munitions stockpile. *Military Operations Research Journal* 12:25-37. *This paper presents a tutorial of Bayesian system reliability assessment for a real problem.*
- T. Graves, M. Hamada, R.M. Klamann, A. Koehler, H.F. Martz. (2008). Using simultaneous higher-level and partial lower-level data in reliability assessments. *Reliability Engineering and System Safety* 93:1273-1279. *This paper shows how to use simultaneous higher-level and partial lower-level data for system reliability assessment. Gromit produces an output of solution (e.g., cut) sets used by YADAS to represent the information in this type of data in terms of component probabilities. This methodology can also be applied to fault trees.*
- C.S. Reese, V. Johnson, M. Hamada and A.G. Wilson. (2005). A hierarchical model for the reliability of an anti-aircraft missile system. Los Alamos National Lab Technical Report LA-UR-05-9281. Submitted to *Journal of the American Statistical Association. This paper considers multi-level lifetime data and the original analysis could have been easily implemented in YADAS. This paper is currently being revised that requires a new application. YADAS has reimplemented the original analysis. It will be easy to adapt the code for the new application.*
- C.M. Anderson-Cook, T.L. Graves, M. Hamada (2008), in preparation. Several new models for systems which are thought to behave almost according to a given structure function (e.g. series or parallel). This approach has been implemented for binomial data at all levels (Reliability Models for Almost-Series and Almost-Parallel Systems, submitted to *Technometrics*), and our work includes a special MCMC algorithm for highly correlated probability parameters. It has also been implemented for arbitrary data forms for the leaf nodes (paper submitted to *Quality and Reliability Engineering International*).
- C. Anderson-Cook, T.L. Graves, and M.S. Hamada (2008), “Resource allocation for reliability of a complex system with aging components,” Los Alamos National Lab Technical Report LA-UR-08-05483. Submitted to *Quality and Reliability Engineering International* Resource allocation in the context of arbitrary data forms for leaf nodes.

2. fault trees

- M. Hamada, H. Martz, C.S. Reese, T. Graves, V. Johnson, A.G. Wilson. (2004). A fully Bayesian approach for combining multilevel failure information in fault tree quantification and optimal follow-on resource allocation. *Reliability Engineering and System Safety*, 86, 297-305. *The simple fault tree considered here did not require YADAS. If a more complicated fault tree had been used, YADAS would have been used. See the next paper.*
- T. Graves, M. Hamada, R.M. Klamann, A. Koehler, H.F. Martz. (2006). A fully Bayesian approach for combining multi-level information in multi-state fault tree quantification. *Reliability Engineering and System Safety* 92:1476-1483. *For multi-state fault trees, the multi-level data combination is not trivial. Gromit produces an output of solution (e.g., cut) sets used by YADAS to represent higher level event data and information in terms of basic event probabilities.*

3. combine random and convenience samples, where production produced small lots

   - T. Graves, M. Hamada, J. Booker, M. Decroix, K. Chilcoat and C. Bowyer. (2007). Estimating a proportion using stratified data arising from both convenience and random samples. Technometrics, 49:164-171. *Small lots required hypergeometric distribution for random samples, extended hypergeometric distribution for convenience samples, and hierarchical model to model quality of lots produced by a common process. This analysis supported a LANL study and set a precedent for using Bayesian methods. YADAS a quick first cut analysis, easily incorporated subsequent model changes, and allowed the use of an alternative algorithm to improve convergence. This paper won the American Society for Quality's Frank Wilcoxon Prize in 2008.*
   - T. Graves and M. Hamada. (2006). Biased reduced sampling: detectability of an attribute and estimation of prevalence. *Quality and Reliability Engineering International*, 22, 385-392. *This paper showed the benefit of biasing the sampling when only small sizes are allowed and how to correct for the biasing in making inferences from such biased samples.*
   - T. Graves (in progress). MCMC Algorithms for Correlated Discrete and Continuous Parameters. *The MCMC algorithms in the lot model problem were highly inefficient, owing to correlations between integer-valued and continuous unknown parameters. We found a more efficient approach that involves treating the integer-valued parameters as discretized versions of latent continuous parameters, and taking advantage of tools for improving efficiency with correlated continuous parameters. This approach may have some potential in image reconstruction problems.*

4. nonlinear regression model where mean is solution to a differential equation

   - Graves, T.L. and S.A. Vander Wiel (2006). Uncertainty Quantification for Hydrolysis Aging in PBX-9501. Los Alamos National Laboratory Controlled Publication LA-CP-06-0942 *The statistical model for the data has mean given by the solution to a differential equation, where some of the parameters in the equation are unknown. This is likely the first Bayesian analysis using such an implicit model. YADAS calls a FORTRAN program that solves a differential equation.*

5. pdf involves an integral

   - S. Michalak, M. Hamada and N. Hengartner. (2006). Bayesian inference for interval-censored data with measurement error, with application to soft error rate estimation. Los Alamos National Lab Technical Report LA-UR-06-7210 (2006). Submitted to *Technometrics. The pdf for Weibull lifetime (in neutrons) data where the data available are interval censored pulse counts, where number of neutrons per pulse is a random variable with known distribution, involves an integral. YADAS allows the calling of functions to* *evaluate expressions; for this paper, the integral is approximated using Laguerre polynomials. Covariates are also incorporated in the modeling.*

6. NDE data comparison with DT data

    - C.S. Reese, P. Deininger, M. Hamada, and R. Krabill. (2008). Exploring the statistical advantages of nondestructive evaluation over destructive testing. *Journal of Quality Technology* 40:259-267. *This paper was motivated by LANL engineers who wanted a statistical argument for nondestructive evaluation (NDE) over destructive testing (DT). This paper considers misclassified pass/fail data (with known misclassification rates) and continuous data with inflated variance due to measurement error for the NDE data. Recently, we reimplemented the analysis and a simulation in YADAS and R for different sample sizes. We also implemented the unknown inflation factor case for the continuous data. We can easily implement the pass/fail data with unknown misclassification rates case. We also want to consider the case where we measure a unit multiple times to see what gains in precision this makes. Again, this requires simple changes to the model which is trivial in YADAS.*

7. T. Graves and S. Vander Wiel (in preparation). Growth over time of features at different angles in a circle, where log slopes are random effects with correlations across angles, and the data are rounded Gammas with standard deviation proportional to square of mean.

8. A. Gelman and J. Hill, in the process of writing their new book on multilevel hierarchical models, published a list of modeling problems at
http://www.stat.columbia.edu/∼cook/movabletype/archives/2006/01/fitting_multile.html
Graves took it as a challenge to develop YADAS code to analyze these six models in one afternoon, and mostly succeeded. Example 3 proved difficult to obtain an efficient algorithm on that time scale, but other software also has this difficulty.

9. Graves, T.L. and R.R. Picard (2003). Seasonal Evolution of Influenza-Related Mortality, Los Alamos National Laboratory Report LA-UR-03-1237. *This paper features a hierarchical model for weekly counts of disease data. We model this with a moving baseline, and a Gaussian seasonal shape with a hierarchical model (in years) for peak timing, width, and integral, and with a further level of hierarchy for different cities.*

10. combining misaligned data

    - *Data collected over time differ when the part is divided into different regions. Regions from different times likely overlap. Multivariate normality and additivity of combined regions is assumed. Alternative covariance structures can easily by entertained. This paper shows how to combine such misaligned data. An interesting planned extension is to misaligned multivariate Poisson counts.*

11. Bayesian variable selection

    - *An implementation of Chipman's proposal for designed experiments was recently implemented in YADAS. Work continues to incorporate new types of covariates corresponding hypotheses of interest in genomic studies which is motivated by Press' LDRD grant. Related covariates will provide a new analysis methodology for industrial experiments. This is work by W. Press, S. vander Wiel, M. Hamada, T. Graves.*

12. early versions of hockey stick model

- *These models supported a LANL study on the impact of lack of data. These model involve a changepoint in which the slope of the line after the changepoint is random. Hierarchical aspects of the problem are incorporated in the models. YADAS was used to quickly implement several models. This is work pursued by S. vander Wiel.*

13. analysis of computer experiments

    - *Standard GASP models have been implemented in YADAS. One of these featured a mixture distribution for the correlation parameters: each of them can be exactly zero or drawn from a Gamma distribution. This application was not pushed very hard: it was only a proof of principle, and at any rate the "Gatt code" exists.*

14. various angular and directional data problems

    - directional data from single population *The data pdf involves an integral known as the Bessel function which can be called from YADAS.*
    - angular rounded data from single population (in progress with C. Anderson-Cook, H. Wu) *The data pdf for a Von Mises distribution involves an integral known as the Bessel function which can be called from YADAS. Also, an approximation of the integral can be used.*
    - angular (and rounded) gauge R&R study, i.e., variance components (in progress with C. Anderson-Cook, H. Wu). Required new update steps for analysis-of-variance-type problems for circular data.
    - The methods above can be applied to rounded and/or ordinal directional data from a gauge R&R study.

15. Preston-Tonks-Wallace model. *Data's expected value are a somewhat complex function of several unknown parameters, and these parameters satisfy a constraint and vary hierarchically according to a multivariate normal distribution with unknown covariance matrix, which we give the Yang and Berger reference prior. The code seems to be "correct" but correlations between parameters render it inefficient.*

16. analysis of nanotechnology experimental data *Paper that C. Anderson-Cook and M. Hamada will discuss with Q. Jia in MST by Jeff Wu and Roshan Joseph and their colleagues involves a multinomial response regression model. We implemented a Bayesian analysis of this model and data easily in YADAS to demonstrate that we can analyze such data. The Bayesian approach also has advantages in optimization by accounting for parameter uncertainty.*

17. T.L. Graves, C.S. Reese, M. Fitzgerald (2003). Hierarchical Models for Permutations: Analysis of Auto Racing Results. Journal of the American Statistical Association, v. 98, no. 462, 282-291. *This paper introduced a powerful family of models for permutation data and illustrated how to do Bayesian computations for these models. Permutation data are of interest in many problems besides sports; for example Google is keenly interested in models for search engine results.*

18. A slightly simpler sports problem is college football or basketball rankings. We routinely publish the results of a Bayesian model with home field advantage at `madison.byu.edu`.

19. YADAS is an environment for research experimentation on MCMC algorithms. See, for example, Graves, T.L., Speckman, P.L., and Sun, D. (2004). Characterizing and Eradicating Autocorrelation In MCMC Algorithms For Linear Models. Los Alamos National Laboratory Report LA-UR-04-0486.

20. YADAS has been used in support of *Bayesian Reliability* by M. Hamada, A.G. Wilson, C.S. Reese, H.F. Martz, Springer (2008). *New YADAS likelihood classes were written to handle Weibull Type II regression data and destructive degradation data. YADAS with R used as a driver solves the reliability planning problems considered in Chapter 9 of this book. The R driver implements a genetic algorithm as well as other optimization algorithms and evaluates a plan by analyzing multiple data sets. Various sample size and planning problems are considered which include binomial and lifetime data sample sizes, an accelerated life test, a degradation data test and resource allocation for system reliability assessment. Many other planning problems could be addressed using these methods.*

21. YADAS has implemented multiple response model from designed experiments with complicated variance-covariance matrix as a complicated function of the covariates. *YADAS employing variance-covariance decompositions using eigenvalue-eigenvector routines.* This methodology can be applied to complicated multiple random effects models for degradation data. The data set exhibits gradual changes in the measurement process, which are estimated nonparametrically. The simultaneous estimation of regression functions and a covariance matrix in a multiple response context is challenging from a model justification and also a model fitting standpoint.

22. Graves, T.L. (2006). The YADAS Resource Allocation Package. Los Alamos National Laboratory Report LA-UR-06-7740. *This report announces a package for performing resource allocation studies, in which the analyst is trying to optimize over an experimental design, and each candidate design needs to be evaluated using one or more MCMC algorithms.*

23. YADAS is enabling research in Bayesian analyses for data with non-standard likelihoods. This effort involves N. Hengartner, B. Williams, T. Graves, S. Reese, M. Hamada, A. Huzurbazar, C. Linkletter, R. Sitter and D. Scott.

24. Random graphs: edges between pairs of nodes are present with probability given by exponential family density. Since likelihood requires evaluation of intractable normalizing constant, we experimented with path sampling to estimate this constant within each MCMC step. In another example, measurements taken from a sample of the nodes were taken, and inferences done on the rest of the graph (including presence and absence of unobserved edges), using a Gaussian model where the correlations between node measurements are determined by the graph structure and where the probability of presence of an edge is a logistic function of geographic distance.

25. Smoothing models: expected value for data is a kernel smooth, or an integral of a linear spline. Both of these were inspired by BioNet.

26. We experimented with an example from Vardeman with an unknown change point.

27. We estimated the states and transition matrix for a simple hidden Markov model.

28. Software reliability: YADAS was used in a few experiments involving novel models for hierarchical software reliability, where the set of possible tests to run can be described using a tree, and the beta-point-mass family is used for level in the hierarchical model.

29. Simulation analysis by R. Picard: covariate x has a target value and an observed value, neither exactly equal to the true value; data y has a cubic polynomial relationship with x, and the parameters of the cubic have a multivariate normal prior.

30. Finite mixture of lognormal distributions or finite mixture of Weibull distributions for lifetime data (complete and censored), where single lognormal or Weibull distribution did not fit very well.

31. Planning for nonlinear degradation model (without random effects), integrated with genetic algorithm and data generation written in R.

32. Planning (resource allocation) for Statistical Sciences paper with binomial system data, component 1 logistic regression model for binomial data, component 2 Weibull distribution for lifetime data, and component 3 linear random effects degradation model for degradation data.

33. Brain imaging: given detector reading, we estimate the location, magnitude, and direction of a desired number of dipoles in a 2D model for a brain.

34. Models for assimilating ratings of subsets of a set of proposals by several raters, allowing for different variances for different raters or for different proposals.

35. Censored failure time data under a mixture of exponentials model, for a high performance computing application.

36. M. Hamada and K.J. Ryan (2008) misclassified ordinal regression data with gauge R&R ordinal data to estimate misclassification probabilities simultaneously.

3 Other Models Which Could Have Been Implemented in YADAS

1. computer and physical experiments with expert opinions

    - C.S. Reese, A.G. Wilson, M. Hamada and H.F. Martz. (2004). Integrated analysis of computer and physical experiments. *Technometrics*, 46, 153-164. Also, A.G. Wilson, C.S. Reese, M. Hamada and H.F. Martz. (2003). Integrated analysis of computer and physical experimental lifetime data. Chapter in *Mathematical Reliability: An Expository Perspective*, Kluwer Academic Publishers, Eds. R. Soyer, T. Mazzuchi and N. Singpurwalla, 183-194. *These papers consider the integration of computer and physical experimental data with expert opinion. The models in both papers were implemented in C but today could have easily been implemented in YADAS and would be if alternative models were considered.*

2. pdf involves an integral

    - T. Graves and M. Hamada. (2005). Making inferences with indirect measurements *Quality Engineering*, 17, 555-559. *This paper considers data whose pdf can be expressed as an integral. A Bayesian version of this paper could easily be implemented using YADAS.*

4 Some generic capabilities of YADAS

YADAS includes components that will be useful in the development of new analyses. Some of them are listed here. Some standard useful densities YADAS can handle are

- Normal, Binomial, Poisson, Gamma, Beta, MultivariateNormalCov (LogCov, Precision), StudentT, Dirichlet, Multinomial, NegativeBinomial, Uniform, Weibull, NegativeLogGamma
- Hypergeometric, ExtendedHypergeometric for biased sampling
- Covariance matrix reference prior, Wishart, InverseWishart
- CensoredLognormal, CensoredWeibull

- For one example, calculated likelihood as function of sufficient statistics for efficiency reasons
- General mixture distributions: currently implemented somewhat nonintuitively, but improvements are in progress.

Some commonly used arguments to density functions are

- Constant, identity, grouped, functional argument (apply arbitrary function to parameters)
- Linear models (and hence generalized)
- Compositions of existing arguments.

Algorithm components include updates for

- Gaussian random walk Metropolis
- On log or logit scale
- Parameters with finite support
- Parameters that sum to one
- Update covariance matrices with Wishart proposals
- Angle parameters
- PermutationParameter
- MultipleParameterUpdates: Add same amount to several parameters; increase some parameters, decrease others; do same things on log or logit scale; emulate matrix reparametrization, others
- Reversible Jump MCMC done in considerable generality. *Used in combination with mixture distributions, this capability makes Bayesian hypothesis testing possible.*

Other features include

- Graves, T. (2005). Automatic Step Size Selection in Random Walk Metropolis Algorithms. Los Alamos National Laboratory Report LA-UR-05-2359. *automatic tuning of Metropolis proposal standard deviations, which YADAS can perform in considerable generality.*
- Automatic algorithm building is an area of considerable potential for YADAS.
- can use importance sampling and/or simulated tempering: sample from "wrong" equilibrium density via MCMC, and reweight
- approximations of parameter-expanded data augmentation have been toyed with.

Some miscellaneous comments related to other interfaces

- We have used D. Temple Lang's SJava package to allow a subset of YADAS models to be defined in R. It works, but is not terribly convenient (debugging in particular is a nightmare).
- We have also defined models in Python (or, more precisely, Jython).

Acknowledgments

Thanks to the YADAS Users' Group, represented by Mike Hamada.